

ウェブコーパスと検索システムを利用した推量副詞とモダリティ形式の遠隔共起抽出と日本語教育への応用

Srdanović Irena^{†,††} · Hodošček Bor^{†,††} ·
Bekeš Andrej^{††} · 仁科喜久子[†]

日本語におけるモダリティ形式および推量副詞と文末モダリティ形式との共起についての体系的な研究は自然言語処理の分野において不十分である。さらに、このような情報は日本語教育の分野においても十分カバーされていない。本稿では、コーパス検索ツール Sketch Engine (SkE) を利用した日本語の推量副詞とモダリティ形式の遠隔共起の抽出を可能にすることとその日本語教育、特に日本語学習辞典への応用の可能性を示すことを目的とする。そのためにまず、複数のコーパスを分析した結果として、モダリティ形式とそのバリエーションの網羅的なリストを作成した。このモダリティ形式は ChaSen でどのように形態素解析されているかを調査し、各モダリティ形式の様々な形態素を新しいモダリティのタグとしてまとめることによって、ChaSen で形態素解析されている JpWaC という大規模ウェブコーパスから抽出した 2 千万語のサンプルヘタグの再付与を行った。最後に、新しくタグ付けされたコーパスをコーパス検索ツール SkE に載せ、「文法関係ファイル」の内容を変更することで、推量副詞と文末モダリティの共起の抽出を可能にした。抽出された共起の結果は 93% 以上の精度で高く評価された。得られた結果は言語資源を利用しての日本語教育への応用の一例として、日本語教育における辞書編集をはじめ様々な教育資源の作成のために、あるいは教室における直接的に利用可能となることを示した。
キーワード：推量副詞，モダリティ，遠隔共起，コーパス検索システム，日本語教育

Extraction of Suppositional Adverb and Clause-Final Modality Form Distant Collocations Using a Web Corpus and Corpus Query System and its Application to Japanese Language Learning

SRDANOVIĆ IRENA^{†,††}, HODOŠČEK BOR^{†,††},
BEKEŠ ANDREJ^{††} and NISHINA KIKUKO[†]

A systematic account of Japanese language modality forms as well as distant collocations between modal adverbs and clause-final modality forms is lacking in the field of natural language processing. The same stands for coverage of this kind of linguistic information in Japanese language education. In order to remedy this deficiency, in this

[†] 東京工業大学, Tokyo Institute of Technology

^{††} リュブリャーナ大学, University of Ljubljana, Slovenia

paper we make extraction of Japanese adverbs and clause-final modality forms collocations possible using the corpus query system Sketch Engine and examine possibilities for its application in Japanese language learning, focusing on learner's dictionaries. First, as a result of analyzing various Japanese language corpora, we create a long list of modality forms and their variations. Then, we examine how ChaSen morphologically analyzes the forms and re-tag a sample of the large-scale Japanese language web corpus, JpWaC, by grouping all morphemes that correspond to individual modality forms together under a new modality tag. Finally, we load the newly tagged corpus into the Sketch Engine (SkE), modify the gramrel file and as a result obtain Word Sketch results for collocations between suppositional adverbs and modality forms. The evaluation of the collocation results shows that the proposed method reaches accuracy of above 93%. The results can be utilized in the creation of Japanese learners' dictionaries or other language material or directly in language teaching or learning.

Key Words: *suppositional adverbs, modality, distant collocations, Sketch Engine, Japanese language learning*

1 はじめに

近年、コロケーションの研究は自然言語処理及びコーパス言語学において盛んになっている。このコロケーションの一種である遠隔共起¹は、頻繁に現れる言語現象であるにも関わらず、研究としては取り上げられていない。日本語における遠隔共起の一つとして推量副詞と文末モダリティ形式²との共起関係があげられ、日本語教育においても重要な問題の一つである。このような共起はモダリティを二重に表現していることにより、テキストにおける重要な語用論的な指標となっている (Bekeš 2008)。Srdanović et al. (2008a) では、工藤 (2000) が示した確率論的性格を有する推量副詞と文末モダリティの「共起」の振る舞いが、複数のコーパスの分析の結果においても確認された。さらに、推量副詞と文末モダリティ形式の共起はテキストの種類によって著しく異なっていることが示され、日本語コーパス資料の分類の可能性が明らかにされた。日本語においてモダリティ形式とそのバリエーションが非常に多いにもかかわらず、そのリストが存在しないため、現在の形態素解析ツールにおいては複数の形態素の連なりからなる様々なモダリティ形式の認識が不可能である。このことから、本研究ではコーパス検索ツール Sketch Engine (SkE) (Srdanović et al. 2008b; Srdanović and Nishina 2008) において日本語の推

¹ 本稿では、あるテキストにおいて、語と他の言語的要素が同時に出現する現象を「共起」と呼ぶ。情報処理およびコーパス言語学において、同じような意味で「コロケーション」という用語も利用される。文末モダリティ形式との共起は語と文法範疇の共起とも考えられ、そのような語と文法範疇の共起は「コリゲーション」という用語を使用することがある。それに対して、語と語の共起は「コロケーション」と呼ばれる (石川 2008)。本章では、両方の現象を「共起」という用語で扱う。「遠隔共起」というのは離れた位置に出現する共起である。

² たとえば、「たぶん最初は発表のスタイルもばらばらで 声もあまり出ないだろう」の例には、「たぶん」は推量副詞、「だろう」は文末モダリティ形式である。推量副詞および文末モダリティは、両方とも、話し手の確信の度合いを表している。

量副詞とモダリティ形式の遠隔共起が検索可能になるように機能の拡張を試みる。実現方法としては、複数のコーパス分析の結果に基づいて、モダリティ形式とそのバリエーションのリストを作成し、ChaSenで認識できるようにした上で、語の文法・共起情報を提示するために推量副詞と文末モダリティ形式との遠隔共起が容易に抽出できるようにする。この抽出結果によって、日本語の学習者、研究者、教師が推量副詞と文末モダリティ形式の共起表現を簡単に調べられ、学習辞書や教科書などの作成に効率的に応用できるようになると考えられる。

2 先行研究

2.1 言語学と日本語教育の分野における研究

南 (1993, 1974) は日本語の文の階層構造を体系的に究明した。文において、接続形式と述部の要素、および接続形式と述部以外の要素の共起制約から、四階層の入れ子構造を確認している。以下の例 (1) では、A 層は述語「いる」1 語、B 層は「この町にも五人ぐらいいはいる」という核の文があり、C 層では副詞「どうやら」と文末モダリティ表現「～らしい」が呼応している。

- (1) { どうやら [この町にも五人ぐらいいは (い) _A] _B らしい } _C (Bekeš 2008)

南 (1993) は、内省ではなく、コーパスにおける共起の頻度を統計的に検証した結果に基づいて、副詞と特定のモダリティ形式の共起の可能性を明確に示している。

工藤 (2000) も、南と同様にコーパスを利用している。副詞と文末モダリティの呼応の関係を検討しながら、それぞれの副詞と特定のモダリティ形式との共起の結びつきの強弱の程度を明らかにし、結びつきの強いものを呼応・共起関係とみなした。推量副詞は対象面から言えば事態実現の確実さ、作用面から言えば話し手の確信の度合いを表している (工藤 2000, p. 204, 5)。この副詞群は「確信」「推測」「推定」「不確定」という四つに区分され、「確信、推測、推定、不確定」の順で、事態実現の確実さおよび話者の確信の度合いが低くなることを連続的な関係で示すものである。推量副詞と共起するモダリティ形式は四つのモダリティタイプ、即ち「に違いない」「はずだ」のような確信 (NEC, necessity), 「だろう」「と思う」のような推測 (EXP, expectation), 「らしい」「みたい」のような推定 (CON, conjecture), 「かもしれない」「かな」のような不確定 (POSS, possibility) に分けられている (工藤 2000, p. 203)。例えば、「多分」は「だろう」「のだろうか」「と思う」「のではないかな」などの推測 (EXP) タイプに属するモダリティ形式と強く呼応する。推量副詞と文末モダリティの呼応の関係は、コーパスにおける共起の統計的な傾向として捉えることができる。工藤のデータについては、3.6 で詳述する。

Bekeš (2008) は、会話コーパスにおける推量副詞と文末モダリティの呼応を分析し、会話における副詞とモダリティの呼応関係の振る舞いを明らかにした。

Srdanović et al. (2008a) および Srdanović et al. (2008) は、推量副詞とモダリティ形式の共起の傾向を複数のコーパスを対象に分析している。コーパスの種類によって推量副詞の分布は異なっており、この副詞と共起する文末モダリティ形式およびモダリティタイプの頻度の傾向にも違いが見られることを明らかにした。また、コロケーション・コリゲーションの研究において語句と語句との共起および語句と文型との共起関係はしばしば研究されるが、遠隔共起関係についての認識はあまりないことから、日本語における推量副詞と文末モダリティ形式の共起を一つの例として取り上げ、遠隔共起関係の重要性を指摘した。

日本語教育の面から、個別の副詞に関して、研究や教育への示唆がいくつかあるが(砂川 2007)、コーパスに基づいた様々なモダリティ形式の整理及び複数のコーパスによる推量副詞とそのモダリティ形式との共起関係の振る舞いについての研究は見られない。また日本語教科書においての扱いも不十分であることから、本研究では、日本語学習辞典の編纂への提案も考慮に入れる。

2.2 自然言語処理における研究

自然言語処理においては、モダリティ形式が機能表現および長単位の研究において部分的に扱われているのが現状であり、日本語のモダリティ形式を中心に扱っている研究はほとんど見られない。先行研究を概観するに際して、推量副詞とモダリティ形式の呼応の研究とモダリティ形式を記載する電子化辞書などの自然言語処理のための言語資源に分け、その研究と自然言語処理の汎用ツールを使用する上での未解決の問題点を述べる。

自然言語処理の汎用ツールを使用する際、意味的にはひとつのまとまりとなっている「かも知れない」「に違いない」などのモダリティ形式が複数の形態素または文節に分割されてしまう問題がある。また、「かもしれない」「かも知れません」「かも」のように、モダリティ形式の表記が文体、丁寧さ、形態素の数などの様々な要素によって異なってくることがある。この場合、モダリティ形式を記載している言語資源を使用すれば、モダリティ形式の自動抽出あるいは認識が可能になる。

松吉 他 (2007) によって編纂された階層構造を用いた機能表現辞書「つつじ」³はその言語資源の一つで、機能表現の一種とも考えられるモダリティ形式がその中にある程度収録されている。また、富士池 他 (2008) は文節を基にした複合辞と連語を含む「長単位」という言語単位を提唱し、「かも知れません」「わけではない」などを1長単位として規定している。しかし、前者は機能表現辞書の作成、後者は「長単位」という形態論の研究であり、いずれもモダリティ形式に特化したものではないため、本稿で扱っているモダリティ形式をすべて含んでいるわけではない。例えば、機能表現辞書「つつじ」には、「といえない」「といえる」「と思う」「と考える」「のか」「ように思う」「気がする」「気がしない」のモダリティ形式が含まれていない。ま

³ <http://kotoba.nuee.nagoya-u.ac.jp/tsutsuji/> から配布されている。

た、「長単位」で認められるのは、助動詞相当として扱われている「かもしれない」「てはいけない」「てはならない」「なければならない」「のだ」「のではない」「わけだ」「わけではない」「わけにはいかない」のみである⁴。木田 他 (2002) は副詞とそれと共に起る表現をまとめて呼応表現と呼び、文中に副詞が現れると、その後に呼応する表現が来ることが予測できることに焦点をあてた。そのため、人手作業による副詞とモダリティの呼応関係を含むタグ付きコーパスを作成し、モダリティ階層関係に基づく呼応関係辞書を編纂した。しかし、人手作業のため、モダリティ形式を自動認識できる言語資源およびシステムは提供していない。

SkE で推量副詞とモダリティ形式の共起関係を抽出するためには、網羅的なモダリティ形式のリストおよび表記のゆれなどを認識できるような言語資源やツールが必要である。そこで、先行研究を参考にしつつ、3.3 節に述べる方法でモダリティ形式およびそのバリエーションのリストを作成し、見出し語と品詞の認定およびその再付与を行うことで、多様なモダリティ形式と推量副詞の関係を SkE 上で検索可能にすることを目指すことにする。

3 SkE での推量副詞とモダリティ遠隔共起の抽出

3.1 SkE とその主な機能

SkE はコーパス検索システムであり、コンコーダンス機能以外に、語に付随する文法とコロケーション情報を表示する機能 Word Sketch や、シソーラス情報や意味的に類似する語の共通点と差異を示す機能 (Thesaurus と Sketch Difference) も備えている。SkE は現在、複数の言語に対応しており (Kilgarriff et al. 2004)、その一つに日本語版がある (Srdanović et al. 2008b)。日本語版は、JpWaC という 4 億語の大規模 Web コーパスを形態素解析ツール ChaSen で解析したデータと、日本語の「文法関係ファイル」から成る。このファイルは、様々な文法関係・共起関係を 22 項目にまとめ、ChaSen の品詞体系を用いた正規表現で定義している。SkE の検索の結果において、Word Sketch 機能は名詞、形容詞、形容動詞、動詞などと共起する要素を表示する。図 1 はその 1 例として、「単語」という語との共起の一部の結果を示している。ここに見られる共起関係は形容詞 (modifier_Ai)、形容動詞 (modifier_Ana)、動詞 (を verb、が verb) などとの共起である。それぞれの欄に表示されている 2 列の数字は、1 列目がコーパスの中の共起頻度を示し、2 列目がその共起の統計的な顕著性 (salience)⁵を示している。表中の 1 列目の数字をクリックすると、例文がコンコーダンスの中で表示される。

SkE に搭載されている大規模ウェブコーパス JpWaC は 4 億語からなり、均衡コーパスを目指して慎重な抽出方法により構築されたものである (Erjavec et al. 2007; Srdanović et al. 2008b)。

⁴ 本稿で扱っているモダリティ形式は 3.3 節を参照されたい。

⁵ 共起関係はダイス係数の統計値に基づいて抽出されている。

modifier Ai 377 5.6	modifier Ana 303 5.0	を verb 3028 4.1	が verb 971 3.2	が Adj 195 2.3
易しい 8 7.85	難解 6 6.86	覚える 215 6.77	聞き取れる 13 7.15	多い 49 3.08
難しい 89 5.61	簡単 66 6.22	並べる 60 6.68	よぎる 5 5.78	難しい 12 2.72
やさしい 9 5.35	適当 9 5.05	区切る 11 5.94	飛び交う 5 5.59	少ない 10 2.42
短い 17 5.04	妙 5 4.95	組み合わせる 16 5.84	浮かぶ 11 4.26	よい 12 1.84
懐かしい 6 5.01	不明 7 4.57	調べる 76 5.53	並ぶ 17 4.12	長い 6 1.77
新しい 63 4.44	便利 8 4.49	思い浮かべる 9 5.48	聞こえる 11 3.59	やすい 8 1.35
長い 27 3.94	特殊 6 4.05	置き換える 12 5.46	出る 154 3.43	ない 17 0.84
近い 9 3.19	適切 6 3.35	ひく 20 5.44	読める 5 3.31	
正しい 8 2.99	重要 20 2.73	読み上げる 7 5.24	見つかる 5 2.79	は verb 426 2.3
やすい 17 2.44	大事 5 2.27	聞き取れる 5 5.02	含む 20 2.75	覚える 17 3.17
にくい 6 2.16	いろいろ 10 2.24	訳す 9 4.92	でる 17 2.75	調べる 8 2.36
少ない 8 2.09	必要 23 1.68	おぼえる 5 4.84	覚える 12 2.65	引く 6 2.29
高い 16 1.94		拾う 12 4.79	現れる 5 2.27	使う 29 1.48
多い 10 0.79		引く 35 4.74	分かる 21 2.25	忘れる 6 1.29

図 1 Word Sketch での「単語」の検索例

最近の研究によると、ウェブコーパスはバランスの取れたデータの傾向を有していることが明らかになっている (Ueyama and Baroni 2005; Sharoff 2006). Srdanović et al. (2008) および Srdanović et al. (2009) では、13 種の日本語のコーパスにおいても推量副詞を分析した結果、JpWaC は、均衡コーパスとして設計された BCCWJ (Balanced Corpus of Contemporary Written Japanese) の中の書籍コーパス (山崎 2006) ともっとも類似した結果を示している。石川 (2009) は BCCWJ と各種のウェブコーパスにおける基本語頻度のデータを比較し、ウェブコーパスは信頼性と有用性を持つことを実証的に示した。SkE の応用の可能性は、日本語学習辞書、日本語学研究、日本語教育などにおいて考えられる (Srdanović and Nishina 2008)。モダリティ形式の多くは形態素の単位が連なった形となっており、ChaSen においては直接認識されていない。そのため、本稿では SkE 日本語版において形態素が連なったモダリティ形式をひとつの単位として抽出可能にし、その結果から日本語教育への応用を検討する。

3.2 推量副詞との遠隔共起の抽出

SkE の日本語版において Word Sketch 機能で副詞と他の語との隣接共起関係が見られる。しかしながら、副詞と離れた場所において共起する語をみるためには、遠隔共起検索機能を追加しなければならない。その方法として、ChaSen の品詞で識別できる要素ならば、SkE の「文法関係ファイル」に単純に追加するだけで Word Sketch で表示できる。ここで副詞と遠隔共起するものとして、1) 副詞と (遠隔) 共起する動詞 (modifies_V), 2) 副詞と (遠隔) 共起する形容詞・形容動詞 (modifies_Adj), 3) 副詞と終助詞 (particle_fin) の三つの共起関係が挙げられる。図 2 は、一例として推量副詞の「たぶん」と共起する最も高頻度の動詞、形容詞、終助詞を示しており、この頻度は遠隔共起も含んでいる。

しかしながら、副詞と文末モダリティ形式との共起については、SkE の設計を変更するだけ

modifies V 2632 8.5	modifies Adj 732	particle fin 201 12.0
思う 47 11.87	大丈夫 10	ね 81 37.33
勘違い 4 9.52	よい 24	よ 61 34.05
想定 4 9.43	いい 30	なあ 13 22.65
想像 6 8.98	無理 11	なあ 8 18.73
違う 18 8.88	すごい 12	な 11 18.23
言う 52 8.45	好き 15	の 10 8.93
にる 12 8.37	強い 14	
なる 141 8.35	本気 6	
ある 116 7.84	怖い 7	

図 2 Word Sketch で「たぶん」との共起の抽出
(終助詞, 形容詞・形容動詞, 動詞の場合)

では, 抽出できない. その理由は, 文末モダリティ形式は形態素解析ツール ChaSen で認識されていないこと, および日本語のモダリティ形式とそのバリエーションのリストが存在しないためである. 次節では, SkE に推量副詞と文末モダリティとの遠隔共起の表示を可能にする方法を検討する.

3.3 モダリティ形式の抽出方法

本論文では, (Srdanović et al. 2008a) の同様の調査に, さらにデータを追加したものを分析する. まず JpWaC コーパスから形態素数 2 千万のサンプルコーパスを作成し, そのコーパスから推量副詞を含む文を抽出し, モダリティ形式がどのように現れているかを調査した. さらにこの調査を基に, モダリティ形式のリストの作成, 複数の形態素に分割されたモダリティ形式および推量副詞のタグの再付与, SkE への適用をする.

SkE 用のコーパスのフォーマットとして, 各単語に対し, 出現形, 見出し語, 品詞の三つのタグを用い, ChaSen の出力結果からそれらのみを使用する. また, 2.2 節で述べたようにモダリティ形式の多くが複数の形態素から構成されるが, SkE では 1 語が ChaSen の 1 形態素に相当するので, モダリティ形式を一つの語にまとめ, コーパスにそれを再付与する必要がある. サンプルコーパスの調査結果から, モダリティ形式の出現には活用・スタイル・表記などの多数のバリエーションがあることが明らかになった. このため, 「だろう」「でしょう」「であろう」「だろ」などの同意のモダリティ形式は一つの代表的なモダリティ形式にまとめ, 見出し語を 31 語に統一した. その 31 項目のモダリティ形式を表 1 に示す.

さらに, モダリティ形式をより網羅的に認識するために, モダリティ形式が他のモダリティ形式やモダリティ形式以外の語と連なって現れる場合も検討した. 下記の文例は, 実際にコーパス中に見られたものである.

目の前の仕事が新鮮にうつることにきっと驚くでしょう. (例 3.3.1)

きっと素晴らしいセミナーをなさることでしょう！（例 3.3.2）

きっと皆さんもそうであろうと思います。（例 3.3.3）

モダリティ形式ではない一定の語および形態素がモダリティ形式に接続する傾向が見られた。例えば、例 3.3.1 のモダリティ形式の「でしょう」に形式名詞「こと」が前接する例として 3.3.2 の「ことでしょう」がある。このように、あるモダリティ形式が、単独で現れる場合と、他の語・形態素と結合して、定型的な組み合わせの語句として現れる場合とで、意味が異なることがある。このことから、モダリティ形式と語・形態素による組み合わせからなる定型的な表現を新しいモダリティ形式として扱うことにした。表 1 に示したモダリティ形式に接続できる 12 個の語群を表 2 に示す。

また、例 3.3.3 のように、モダリティ形式が連続して現れる傾向が明らかになった。あるモダリティ形式が、単独で現れる場合と、他のモダリティ形式と接続して現れる場合があり、その語句の出現の様相で意味が異なることがある。本稿では、例 3.3.3 の「であろうと思います」のようなモダリティ形式を複合的なモダリティ形式、「であろう」のようなモダリティ形式を単独モダリティ形式と呼ぶことにする。SkE では、複合的なモダリティ形式を別々に扱うのではなく、一つの新しいモダリティ形式としてまとめて扱うことにした。例 3.3.3 のモダリティ形式は「だろうと思う」になり、連続している形態素をまとめて SkE の 1 単語とする。

表 3 と表 4 からわかるように、同じ意味・機能の表現であるにもかかわらず、ChaSen による見出し語が漢字とひらがなで区別されてしまう場合や、品詞タグが異なっている場合がある。これはモダリティ形式に限らない。「必ずしも」や「ひょっとしたら」などの複数の形態素から構成される副詞も同じ問題を含んでいる。そこで、すべてのモダリティ形式と推量副詞が見出し語と品詞のレベルで統一されるようにコーパス中のモダリティ形式および推量副詞の見出し

表 1 JpWaC サンプルコーパスに出現するモダリティ形式の見出し語

かしら、かな、かもしれない、ざるを得ない、そうだ、だろう、てはならない、といえない／といえる、とは限らない、と思う、と考える、なくてはいけない、なくてはならない、なければいけない、なければならぬ、に違いない、のか、のだ／のではない、はずだ、べきだ、みたいだ、ようだ、ように思う、らしい、わけだ／わけではない、わけにはいかない、気がする／気がしない

表 2 モダリティ形式と接続可能な語・形態素の見出し

こと、もの、という、の、か、かな、な*、は*、も*、なの*、だ、ではない

*「は」が「とはいえない」に現れるように、モダリティ形式に挿入する語・形態素

表 3 JpWaC コーパスにおいて ChaSen による見出し語が異なる場合の処理

処理前			処理後		
出現形	見出し語	品詞	出現形	見出し語	品詞
おもう	おもう	動詞—自立	おもう	思う	動詞—自立
思う	思う		思う		

表 4 JpWaC コーパスにおいて ChaSen による品詞タグが異なる場合の処理

処理前			処理後		
副詞—一般出現形	見出し語	品詞	出現形	見出し語	品詞
必ず	必ず	副詞—助詞類接続	必ず	かならず	副詞—一般
かならず	かならず		かならず		

表 5 副詞とモダリティ形式の関係を定める規則

*DUAL =Adv/modality
2:"Adv.*" "P.advzer"? []* 1:"Mod."

語と品詞を再付与した⁶。ここで、モダリティ形式の品詞タグはすべて「Mod」という新しいカテゴリに書き換えた。これにより、モダリティ形式と推量副詞の表記が文章のスタイルおよび ChaSen の誤解析によって異なる場合でも、網羅的にそれらを SkE で検索したり、違いを見たりすることが可能になる。

以上の方法によって得られたモダリティ形式の数は、代表的なモダリティ形式は 31 個、組み合わせで現れるモダリティ形式は 596 個、その出現形のバリエーションは 2,641 個である。

以上の点を考慮し、抽出する様々な共起関係を定める SkE の「文法関係ファイル」に表 5 の副詞とモダリティ形式の関係を定める規則を追加する。

この規則は、SkE において実装されている正規表現と品詞及び単語に基づいた、「corpus query syntax (コーパス検索シンタクス)」を適用している。「DUAL」は副詞とモダリティ形式の関係を両方向から検索できる規則である。「Adv」と「modality」はそれぞれ「副詞」と「モダリティ形式」を意味する。この規則では、副詞に続いて最初に現れるモダリティ形式のみとの関係を抽出することではなく、文末まで現れる他のモダリティ形式との共起が多くあることから、すべてのモダリティ形式との可能性をみることにした。また、この規則で得られる推量副詞とモダリティ形式の共起は構文を考慮していない。そこで係り受け解析器 CaboCha を使用した共起の抽出方法も考えられるが、本手法と比べていくつかの利用上の欠点がある。一つは、モダリティ形式が係り受けの単位となる文節の境界を越えることがある。他の一つは、CaboCha が

⁶ そのために、SkE で検索する際に、統一した表現で検索する必要がある。例えば、すべての副詞は漢字ではなく、ひらがなで記入して検索することである (多分→たぶん)。

学習したドメインとは異なる JpWaC のようなウェブ文章における遠隔共起の正確な係り受け解析が難しいと考えられることである。

3.4 抽出結果

JpWaC の 2 千万語サンプルコーパス中に見られる複数のモダリティ形式とそのバリエーションにモダリティタグを付け、新しい共起関係を SkE の「文法関係ファイル」に組み込んだことで、推量副詞と文末モダリティの共起が SkE の Word Sketch の機能の中で表示できるようになった⁷。

例として、「たぶん」「どうやら」「もしかしたら」「かならずしも」とそのモダリティ形式との共起の表示結果を図 3 に示す。それぞれの推量副詞が多様なモダリティ形式と共起していること、結びつくモダリティ形式のうち代表的なものがそれぞれの副詞で異なっていることが分かる。「たぶん」は「のだろう」「だろう」「と思う」「のだと思う」などの推測 (EXP) のモダリティタイプとよく共起する傾向がある。「どうやら」は推定 (CON) の「らしい」「ようだ」「のようだ」「みたいだ」など、「もしかしたら」は不確定 (POSS) の「かもしれない」「のかもしれない」「のかな」など、「必ずしも」は確信 (NEC) の「とはかぎらない」「わけではない」「ものではない」などとよく共起する傾向がある。

表 6 では、コーパスにおける推量副詞の分布とそれぞれの副詞と顕著性計指数が高い項目の順に 5 つのモダリティ形式が並んでいる。

まず、推量副詞の分布から見ると、高頻度の副詞は「たぶん」「かならず」「おそらく」「きっと」「ぜったい」などである。推量副詞とモダリティ形式の共起関係を見ると、「のだろう」「だ

たぶん		どうやら		もしかしたら		かならずしも	
modality	959 18.2	modality	382 20.3	modality	269 20.7	modality	249 18.1
のだろう	124 36.59	らしい	121 48.54	かもしれない	102 43.36	とは限らない	53 51.05
と思う	221 35.87	ようだ	95 40.61	のかもしれない	59 39.67	わけではない	35 34.77
だろう	175 35.23	のようだ	27 31.21	のかな	10 19.06	ものではない	11 21.89
のだと思う	32 30.22	みたいだ	16 26.2	のではない	8 15.87	ことではない	2 21.09
のではないかと	28 24.82	そうだ	21 21.46	か	5 12.04	といえない	8 20.93
と思うのだ	24 22.54	らしいのだ	5 16.36	のでもない	5 12.04	ではないと思う	4 16.02
ことだろう	15 18.74	ことらしい	3 13.33	のでもないかと	5 12.04	ものではないと思う	3 15.83
はず	22 18.64	のだ	24 12.04	に違いない	3 10.41	と思う	17 14.12
のだ	73 17.3	と考える	4 9.39	と思う	11 10.3	と考える	6 13.69
ではないだろう	9 16.82	のだろうか	5 8.77	のか	2 10.15	のだ	23 13.66
気がする	16 16.39	のかもしれない	4 8.53	のだ	13 8.87	ようだ	8 13.06
のではない	13 16.07	かもしれない	2 8.06	だろう	6 7.36	気がする	5 10.85
のかもしれない	13 14.97	のではないかと	3 7.59	気かする	3 7.21	ではないのだ	2 10.74
かな	13 14.88	と思う	8 6.98	のだろうか	3 6.47	かもしれない	6 10.47
かもしれない	18 14.78	かな	3 6.73	のだ	3 6.47	だろう	2 10.31
に違いない	2 14.34	のだろう	4 6.36	らしい	4 9.34	のではない	4 10.19
そうだ	13 12.99	のか	4 4.4				
のではないだろう	8 12.77						

図 3 Word Sketch における推量副詞とモダリティ形式の共起表現検索結果
 (「たぶん」、「どうやら」、「もしかしたら」、「かならずしも」とモダリティ形式との例)

⁷ 現時点で通常の使用では出来ないが、ユーザからの要請により、SkE のウェブページで JpWaC のサンプルコーパスにおいて副詞とモダリティの共起を検索することができる。

表 6 JpWaC のサンプルにおける高頻度の推量副詞と文末モダリティ形式の共起関係

タイプ	副詞	頻度	上位の 5 つのモダリティ形式 (共起頻度、共起顕著性において)															
推測	たぶん	1527	のだろう	124	36.5	と思う	221	35.87	だろう	175	35.23	のだと思う	32	30.22	のではないかと	28	24.82	
確信・推測	かならず	1448	はず	31	28.09	のだ	95	27.68	と思う	41	21.96	だろう	20	16.29	わけだ	6	11.57	
推測	おそらく	1341	だろう	269	41.2	のだろう	128	36.7	ことだろう	37	29.06	と思う	131	28.38	に違いない	21	26.0	
推測・確信	きっと	1340	のだろう	160	41.04	だろう	159	34.54	に違いない	34	33.09	ことだろう	44	32.13	はず	52	29.05	
確信・推測	ぜったい	1294	のだ	87	25.35	だろう	40	22.56	と思う	41	20.92	はず	17	20.12	べき	13	16.66	
推定	どうも	1022	ようだ	59	34.4	らしい	44	33.65	気がする	28	27.32	のだ	63	22.16	のようだ	12	21.52	
確信	ぜったいに	816	てはならない	21	34.94	のだ	79	26.43	べき	15	19.49	と思う	29	19.08	はず	12	17.88	
推定	どうやら	548	らしい	121	48.54	ようだ	95	40.6	のようだ	27	31.21	みたいだ	16	26.2	そうだ	21	21.46	
確信・推測	たいてい	497	のだ	45	23.71	と思う	17	16.79	だろう	14	16.36	はず	6	13.51	ように思う	3	9.99	
推測	よほど	409	のだろう	34	28.12	だろう	22	18.87	と思う	25	18.55	のだ	31	17.26	のか	17	16.45	
確信	かならずしも	382	とは限らない	53	51.05	わけではない	35	34.77	ものではない	11	21.89	ことではない	9	21.09	といえない	8	20.93	
不確定	もしかしたら	316	かもしれない	102	43.36	のかもしれない	59	39.67	のかな	10	19.06	のではないかと	8	15.87	のではないかと	5	12.04	
確信・推測	あんがい	187	のかもしれない	11	22.14	気がする	9	19.42	のだ	18	16.13	かもしれない	8	16.1	だろうかと思う	2	15.53	
不確定	ひょっとしたら	81	かもしれない	25	30.48	のかもしれない	17	26.67	と考えるのかもしれない	1	10.25	のではないかと	2	8.16	かもしれないのだ	1	7.87	
確信・推測	たいがい	72	のだ	6	11.16	と思う	4	9.93	わけだ	2	8.8	のかなと思う	1	8.71	だろう	3	8.69	
推測	さぞ	31	ことだろう	6	20.5	だろう	8	16.39	のだろう	5	14.41	だろうと思う	2	11.59	だろうと考える	1	9.16	
推測 (推定)	おおかた	24	ものと思う	1	8.46	ものだろう	1	7.77	と思う	2	7.42	のかな	1	6.72	と考える	1	6.52	
	ことによると	2	らしい	1	7.46	べき	1	7.02										

ろう」「と思う」などの「推測」を表すモダリティ形式および「はず」「のだ」「に違いない」などの「確信」を表すモダリティ形式がもっとも高頻度の共起関係として出現することがわかる。ある副詞が「推測」のモダリティ形式と共起する傾向があれば、その副詞の共起の中では、「確信」のモダリティ形式もよく見られる。たとえば、「きっと」と「おそらく」と共起する「に違いない」などである。逆に、副詞が「確信」のモダリティ形式と共起する場合、その副詞は「推測」のモダリティ形式とも共起する。たとえば、「かならず」「ぜったい」と、「だろう」「と思う」という「推測」の文末形式が共起する場合である。

JpWaC のサンプルデータと均衡コーパスである BCCWJ 中の書籍のコーパスにおける推量副詞の分布と推量副詞と文末モダリティの共起を比較した結果、類似した分布及び共起関係の傾向を示していることが明らかになった (Srdanović et al. 2009)。

SkE 中にある Word Sketch 機能だけでなく、コンコーダンス、Sketch Difference、シソーラスなどの機能においても推量副詞とモダリティ形式について検索できる。たとえば、類似している二つの語の間の共起の傾向における共通点と差異が閲覧できる Sketch Difference という機能では、「たぶん」と「きっと」を比較する際、「きっと」は「のだろう」「に違いない」「ことだろう」「はず」との共起が多く、確信のモダリティタイプに近い性質が見られる。一方、「たぶん」のほうは推測を表す「と思う」「のだと思う」「のではないかと」「と思うのだ」とよく共起することから、「きっと」より推量の度合いが低いと考えられる。

3.5 SkE で抽出された推量副詞と文末モダリティ形式の共起の評価

本節では、Word Sketch による推量副詞と文末モダリティ形式との共起の判定の精度を評価する。推量副詞と共起する代表的なモダリティタイプが四つあり (工藤 2000; Srdanović et al. 2008a), それぞれのタイプから一つの高頻度の副詞を評価のために選択する。また、それぞれの副詞と共起するモダリティの検索結果から、四つのモダリティ形式を選び (Word Sketch に表示するモダリティ形式のリストの最初から四番目までの四項目)、コンコーダンス機能から選

表 7 評価の結果

選択肢	回答		
	評価者 A	評価者 B	評価者 C
(正) 副詞とモダリティ形式の共起は正しい	93.1%	93.8%	96.9%
(?) 副詞とモダリティ形式の共起は部分的に正しい	1.9%	1.2%	1.9%
(誤) 副詞とモダリティ形式の共起は正しくない	5.0%	5.0%	1.2%

んだこの副詞と共起するモダリティ形式の 10 例文をランダムに取り出す。従って、合計 160 の例文を評価することとする。

選択した例文について日本語教育専門家二名 (A,B), 自然言語処理専門家一名 (C) から評価を受ける。選択肢は「(正) 副詞とモダリティ形式の共起は正しい」, 「(?) 副詞とモダリティ形式の共起は部分的に正しい」, 「(誤) 副詞とモダリティ形式の共起は正しくない」であり、コメントを記入できる欄もある。

表 7 に見られるように「(正) 副詞とモダリティ形式の共起は正しい」の回答から、精度は高く評価されていることがわかる。正しく評価されていない例文と評価者のコメントを見ると以下のようなことがわかる。

- 1) モダリティ部分とみなす部分に誤解析となる例文がある。
- 2) 日本語として正しくない例文がある。
- 3) コーパスには重複する例文がある。
- 4) 副詞とモダリティ形式が共起していない例文がある。
- 5) 表示されているものとは異なる他のモダリティ形式と共起している例文がある。

1), 2), 3) は比較的修正が容易である。1) については、部分的に抽出したモダリティ形式をバリエーションリストに追加すること、タグの付け方の改善で解決できる。また、2), 3) については、正しくない文および重複している文の排除を行うことで処理できる。4) と 5) の原因は、形態素解析の誤りに起因する複数の文がコーパスに入っている場合と、構文が複雑なため副詞が自動的に指定されたモダリティ形式と共起しない場合が考えられる。今後の課題として、構文中の呼応関係を示すための構文解析プログラムを改善する必要がある。

3.6 工藤データとの比較

前節では共起項目検索結果の精度が高いと評価されたことを述べたが、さらにこの結果を日本語研究における推量副詞と文末モダリティ形式の呼応関係についてよく精査された研究 (工藤 2000) と比較する。工藤によると副詞とモダリティ形式との呼応は程度の問題だと考えられ、それぞれの副詞は特定のモダリティタイプに属するモダリティ形式とよく共起する傾向があるとしている (図 4)。たとえば、「きっと」は確信のモダリティタイプに属する「する・のだ」「に

副詞／述語	確信				推測			推定				不確定				計	推量以外の用法
	する・のだ	にちがいない	に決まっている	はずだ	だろう・まい	と思われる	のではないだろうか	らしい	と見える	ようだ・みたいだ	しそうだ	かもしれない	だろうか	せぬとも限らぬ	する節がある		
きっと	139	38	8	3	66	12				1	4	8				279	85
かならず	17	5	2	1	11											36	146
ぜったい(に)	48															48	38
おそらく	31	18		1	112	5	10	2		1		2				182	--
たぶん	19	1		2	74		1	1			2	3				103	--
さぞ					52		1				1					54	--
おおかた	2	1			24		1									28	13
たいてい	3			1	7											11	80
たいがい	2				4											6	33
どうやら	5						1	29		10					1	46	39
どうも	13	1					6	24				1				45	385
よほど	6	2			7		2	12	9	3			2			43	150
あるいは					3	2	4					53	3	1		66	69
もしかすれば	2			1	1	1	11					30				46	--
ひょっとしたら	2						7					16	1			26	--
ことによると	1						4					7	1	1		14	--
あんがい		1			1		3	1			1	8				15	81

図 4 工藤による推量副詞と文末モダリティのコーパス分析結果

「違うない」などの形式と最も共起する。「おそらく」は推測の「だろう・まい」「と思われる」「のではないだろうか」と多く共起する。「どうやら」は確定の「らしい」など、「もしかしたら」は不確定の「かもしれない」などと多く共起する傾向が見られる。

本研究で抽出されたデータにも、工藤のデータと同様に推量副詞と共起する四つのモダリティタイプが見られる。このデータにも、推測と確信のモダリティタイプとの共起は最も多く見られる。工藤のデータに見られる様々なモダリティ形式は JpWaC にも表れる。しかし副詞の頻度の順序、およびそれぞれの副詞の共起傾向には差異も見られる。「きっと」との共起は確信のモダリティタイプより、推測のモダリティタイプの方が多く現れる。工藤のデータにはやや古い用語も現れ、例えば、「さぞ」の頻度は、JpWaC より頻度がかかなり高い。この差異の理由は、Srdanović et al. (2008) に複数のコーパスの分類結果で指摘されているように、コーパスの種類が異なるためである。

4 日本語教育への応用

SkE は既に多数の言語において語学教育、辞書作成などに利用されており (Smrž 2004; Smith et al. 2007; Kilgarrieff and Rundell 2002), 日本語版も日本語教育などのために役立つシステムになると考えられる (Srdanović and Nishina 2008). 本研究で抽出できるようになった推量副詞と

モダリティ形式の共起に関する結果は、日本語学習のシラバス、教科書、学習辞典などの学習資源を作成するために利用できる。また、抽出された共起関係は学習者や教師が授業中に参照したり、授業の準備をしたりするためにも利用できる。さらに、整理したモダリティ形式とモダリティ形式のバリエーションのリストが日本語教育のための言語処理の資源など様々の目的のために利用できる。

検索方法の面からみると SkE で推量副詞とモダリティ形式の共起に関して、いくつかのメリットが見られる。(1) ウェブ上で推量副詞と文末モダリティの共起が瞬時に効率的に抽出できる。(2) 共起に関しての高頻度の情報と統計的に重要な共起情報が得られ、学習項目に優先順位を付けることができる。(3) Sketch Difference の機能において、類似している副詞の共起関係の傾向における差異を簡単に把握できる。(4) コンコーダンス機能で、例文を細かく考察でき、複雑な遠隔共起も検索でき、様々なコンコーダンスオプションも利用できる。

一つの応用の例として、日本語の学習辞典の編集・改善案を具体的に検討する。抽出できるようになった推量副詞と文末モダリティ形式の共起関係の結果を 3 種の日本語学習辞典と比較する。学習者によく利用されている和英辞典 2 種 (『ジーニアス和英辞典』、『ニューセンチュリー和英辞典』) と、中級レベル以上の日本語学習者にとって問題となる文型を網羅的に集めた辞典 (『日本語文型辞典』) を対象とする。表 8 は、これらの辞典においてどの推量副詞の項目がカバーされているか、また記載されている各副詞の項目の中で、その例文中にどのモダリティ形式が現れているかを示している。

推量副詞の分布から見ると、和英辞典 2 種には対象となるすべての推量副詞が現われている。「文型辞典」には、現われない副詞もある。「あんがい」「おおかた」「ことによる」のような最も低頻度の副詞以外に、コーパス中にかなり出現している「ぜったい」「ぜったいに」が現われていない。本研究で抽出した結果によると、この二つの副詞はかなり出現することから、中級以上の日本語学習辞典にも含めることを考慮する必要がある⁸。反対に、「さぞ」「たいがい」は JpWaC のコーパスにはあまり現れない。特に「さぞ」はやや古い表現だと考えられる⁹。

推量副詞と共起するモダリティ形式については、和英辞典 2 種の例文中にモダリティ形式が多く現れている。しかし、それぞれの辞典で例文に現れるモダリティ形式は異なっており、特定の副詞と共起する代表的なモダリティ形式が足りない場合もかなりある。たとえば、「たぶん」には、「かもしれない」「ものだ」の代わりに、「のだろう」「と思う」を入れた方がよい。また、「おそらく」には、「と思う」「に違いない」も追加したほうが良いと考えられる。「文型辞典」では、推量副詞のすべての項目に網羅的にモダリティ形式が現れていることが評価できる。しかし、この辞典においても、コーパス中に高い頻度および高い顕著性で現れるモダリティ形

⁸ Srdanović et al. (2009) の研究によると、この副詞はインフォーマルの会話コーパスでは非常に高頻度で、均衡コーパスの書籍と JpWaC、新聞コーパス、知恵袋のウェブコーパス、フォーマルな会話コーパスでもかなり頻度が高い。

⁹ この副詞は教科書系のコーパスでは頻度が非常に高いが、均衡コーパスの書籍では、頻度が非常に低い。

表 8 日本語辞書における推量副詞と文末モダリティ形式との共起

副詞	ジーニアス	ニューセンチュリー	文型辞典
あるいは	▲かもしれない	▲かもしれない	▲かもしれない ●のだろう、と思われる
あんがい	△	▲のかもしれない	/
おそらく	▲でしょう、のだろう、だろう	△	▲だろう、ものと思われる、にちがいない
おおかた	△	△	/
かならず	▲でしょう、なさい、だろう	▲でしょう、てください、必要がある、ことにしている	▲てください、しなければならない、ようにしよう
かならずしも	▲とは限らない、というわけではない、ものではない	▲とは限らない、とはいえない	▲とは限らない、ではないと私は思っている、しなければならないものだと思っているわけではない ●わけではない、とはかぎらない
きっと	▲のです、にちがいない、んだ、だろう	▲だろう、に違いない、でしょう	▲でしょう、だろう、にちがいない、てください
ことによると	△	▲かもしれない	/
さぞ(かし)	▲でしょう	▲でしょう、ことでしょう	▲ことだろう、ことございましょう
たいがい	▲だろう	△	(ことにしている、ことになっている)
たいてい	▲だろう	▲であろう	「たいがい」と同じ例文・説明
たぶん	▲だろう、かもしれない、でしょう	▲でしょう、だろう、かもしれない、ものだ	▲でしょう、ておこう、でしょうか、と思う
どうも	▲ようだ、ように思う	▲そうだ、ようだ、んだ、気がしない	▲そうだ、ようだ、らしい
もしかしたら	▲かもしれない	▲かもしれない	▲かもしれない、のではないだろうか
ひょっとしたら	▲だろう、かもしれない、でしょうか	▲のかもしれない、かと思う	/
よほど	△	▲ようだ	▲に違いない、らしい、んだろう、んだ、んだらうと思う
ぜったい	△	△	/
ぜったいに	▲してはいけない、するな	△(不可能だ、ことはない)	/

△副詞は辞書にあるが、共起するモダリティ形式は記載されていない

▲モダリティ形式が例文にある

●モダリティ形式が他の表現として言及されている

式のデータを導入することについて検討の余地がある。たとえば、「おそらく」という副詞に関し、「ものと思われる」のかわりに、「のだろう」「ことだろう」「と思う」のモダリティ形式を載せたほうがよい。「かならず」の項目では推量のモダリティ形式が現われていないので、「はずだ」「のだ」「だろう」「に違いない」を追加した方がよい¹⁰。

以上の指摘は、均衡コーパスに基づく調査結果によるものであり、一般的な日本語習得を目的とした辞典の作成のためのものである。しかし、コーパスの種類によって代表的なモダリティ形式が異なるので(Srdanović et al. 2008a, 2008), 学習辞典にもこの調査結果を学習目的に合わせて反映させることが望ましい。一つの例として、理系教科書、自然言語処理の論文集のよう

¹⁰ JpWaC における最も高頻度の副詞とモダリティ形式の共起関係は均衡コーパスの書籍と類似した結果が見られる。

な理系の特定コーパスの調査において、均衡コーパスとは異なり、「必ずしも」が他の推量副詞より頻度が非常に高いことがわかった。このような結果を反映させることで、辞典に特定目的の学習をカバーすることが可能となる。

5 まとめと今後の課題

本稿では、推量副詞と文末モダリティに関連する先行研究を調査した結果から、代表的なモダリティ形式に関する実証的な情報と網羅的に作成されたモダリティ形式のリストが欠如していることを明らかにした。次に、コーパス検索システム SkE の主な機能を紹介し、SkE に推量副詞と文末モダリティ形式などの遠隔共起関係の機能の追加を、次の 1) ～4) までの手順で行った。

- 1) モダリティ形式とそのバリエーションのリストの作成,
- 2) ChaSen によるモダリティ形式の形態素解析の検定とモダリティ形式のタグの作成,
- 3) ChaSen で形態素解析されたコーパスへのタグの再付与 (モダリティ形式のタグの追加),
- 4) 「文法関係ファイル」に副詞とモダリティ形式の共起関係を定める規則の追加.

次に、SkE 中の Word Sketch で抽出が可能になった推量副詞と文末モダリティ形式の共起の評価を行い、抽出された共起項目が高く評価されたことを示した。さらに、得られた結果を推量副詞についての工藤の日本語研究と比較し、共通点と差異を明らかにした。最後に、SkE の Word Sketch 機能で推量副詞とモダリティ形式の遠隔共起をどのように日本語教育に応用できるかを検討した。具体的な例として、日本語学習辞典における推量副詞と文末モダリティ形式の共起の扱いを検討した。

今後の課題として、モダリティ形式とそのバリエーションのリストを xml 形式で作成、ウェブコーパス以外のデータに基づいてさらに改善することを目指す。また、様々なコーパスを利用することで、推量副詞と文末モダリティ形式の共起のデータをジャンル別に比較することで、日本語教育における目的別の学習への活用を考える。

参考文献

- Bekeš, A. (2008). “Japanese suppositional adverbs: probability and structure in speaker-hearer interaction.” *Linguistica*, 48, pp. 277–292.
- Erjavec, T., Srdanović, I., and Kilgarriř, A. (2007). “A large public-access Japanese corpus and its query tool.” In *CoJaS 2007*.
- 富士池優美 他 (2008). 『現代日本語書き言葉均衡コーパス』における長単位の概要. 科学研究費補助金特定領域研究「日本語コーパス」平成 19 年度公開ワークショップ (研究成果報告会) 予稿集, pp. 51–58.

- 石川慎一郎 (2008). 英語コーパスと言語教育. 大修館書店.
- 石川慎一郎 (2009). 日本語基礎研究における非統制型・統制型・媒介型 Web as Corpus の可能性—言語コーパスにおける基本語頻度の安定性について. 特定領域研究「日本語コーパス」平成 20 年度公開ワークショップサテライトセッション予稿集, pp. 29–38.
- 木田敦子, 高梨克也, 乾裕子, 井佐原均 (2002). 文構造の漸進的予測を可能にする日本語の諸特徴の分析. 言語処理学会第 8 回年次大会発表論文集, pp. 65–68.
- Kilgarrieff, A. and Rundell, M. (2002). “Lexical Profiling Software and its Lexicographic Applications—a Case Study.” In *EURALEX 2002 Proceedings*, pp. 807–818.
- Kilgarrieff, A., Rychly, P., Smrž, P., and Tugwell, D. (2004). “The Sketch Engine.” In *Proc. Euralex*, pp. 105–116.
- 工藤浩 (2000). 副詞と文の陳述のタイプ. 森山卓郎, 仁田義雄, 工藤浩 (編), 日本語の文法 3 モダリティ, pp. 161–234. 岩波書店.
- 松吉俊, 佐藤理史 (2007). 体系的機能表現辞書に基づく日本語機能表現の言い換え. 言語処理学会第 13 回年次大会発表論文集, pp. 899–902.
- 南不二男 (1974). 現代日本語の構造. 大修館書店.
- 南不二男 (1993). 現代日本語文法の輪郭. 大修館書店.
- Sharoff, S. (2006). “Open-source corpora: using the net to fish for linguistic data.” *International Journal of Corpus Linguistics*, **11** (4), pp. 435–462.
- Smith, S., Chen, A., and Kilgarrieff, A. (2007). “A corpus query tool for SLA: learning Mandarin with the help of Sketch Engine.” In *Practical Applications in Language and Computers—PALC 2007*.
- Smrž, P. (2004). “Integrating Natural Language Processing into E-learning—A Case of Czech.” In *Proceedings of the Workshop on eLearning for Computational Linguistics and Computational Linguistics for eLearning, COLING 2004*, pp. 106–111.
- Srdanović, I., Bekeš, A., and Nishina, K. (2008a). “Distant Collocations between Suppositional Adverbs and Clause-Final Modality Forms in Japanese Language Corpora.” In Tokunaga, T. and Ortega, A. (Eds.), *LKR 2008, LNAI 4938*, pp. 252–266. Springer-Verlag Berlin Heidelberg.
- Srdanović, I., Erjavec, T., and Kilgarrieff, A. (2008b). “A web corpus and word-sketches for Japanese.” 自然言語処理, **15** (2), pp. 137–159.
- Srdanović, I. and Nishina, K. (2008). コーパス検索システム Sketch Engine の日本語版とその利用方法. 日本語科学, **23**, pp. 59–80.
- Srdanović, I., Bekeš, A., and Nishina, K. (2008). 複数のコーパスに見られる副詞と文末モダリティの遠隔共起関係. 科学研究費補助金特定領域研究, 「日本語コーパス」平成 19 年度公

開ワークショップ（研究成果発表会）予稿集, pp. 223–230.

Srdanović, I., Bekeš, A., and Nishina, K. (2009). BCCWJにおける推量副詞とモダリティ形式の共起. 科学研究費補助金特定領域研究「日本語コーパス」平成 20 年度公開ワークショップ（研究成果報告会）予稿集, pp. 237–244.

砂川有里子 (2007). 機能語のコロケーション (1) 副詞との共起について. テクニカル・レポート, 科学研究費補助金特定領域研究日本語教育班第 2 回公開会議.

Ueyama, M. and Baroni, M. (2005). “Automated construction and evaluation of a Japanese web-based reference corpus.” In *Proceedings of Corpus Linguistics*.

山崎誠 (2006). 代表性を有する現代日本語書き言葉コーパスの設計. 国立国語研究所 (2006) 所載, pp. 63–70.

略歴

Srdanović Irena: 1997 年ベオグラード大学文学部日本語学科卒業. 2006 年リュブリャーナ大学文学部一般言語学・比較言語学研究科修士課程修了. 2009 年 9 月東京工業大学大学院社会理工学研究科人間行動システム専攻博士課程終了. 博士（学術）. 言語処理学会, 日本語教育学会会員.

Hodošček Bor: 2007 年米国メリーランド大学カレッジパーク校人文学部日本語学学科卒業. 同年リュブリャーナ大学文学部一般言語学・比較言語学研究科修士課程入学. 2008 年東京工業大学大学院社会理工学研究科人間行動システム専攻修士課程入学. 現在に至る. 言語処理学会会員

Bekeš Andrej: 1971 年リュブリャーナ大学理学部数学学科卒業. 1975 年大阪大学大学院理学研究科数学専攻修士課程修了. 1981 年筑波大学大学院人文社会科学部研究科文芸・言語専攻入学. 1986 年同課程修了. 1988 年リュブリャーナ大学社会学部専任講師. 2002 年同大学文学部正教授（日本学）, 現在に至る. 文学博士. 日本語教育学会会員.

仁科喜久子: 1969 年東京女子大学・文理学部卒業. 1977 年同大学大学院文学研究科修士課程修了. 1986 年埼玉大学専任講師, 1988 年東京工業大学助教授を経て, 1996 年同大学留学生センター教授（社会理工学研究科兼任）, 現在に至る. 博士（学術）. 言語処理学会, 日本語教育学会, 教育工学会会員.

（2008 年 8 月 31 日 受付）

（2009 年 4 月 15 日 再受付）

（2009 年 6 月 1 日 再々受付）

（2009 年 7 月 17 日 採録）